RGB-D Face Recognition via Deep Complementary and Common Feature Learning

Hao Zhang[1,2], Hu Han[*,1], Jiyun Cui[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]*Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*CAS Center for Excellence in Brain Science and Intelligence Technology*
{*hao.zhang,jiyun.cui*}*@vipl.ict.ac.cn;* {*hanhu, sgshan, xlchen@ict*}*@ict.ac.cn*

***Abstract*—RGB-D face recognition has attracted increasing attentions in recent years because of its robustness in unconstrained environment. However, existing approaches either handle individual modalities using completely separate pipelines or treat all the modalities equally using the same pipeline. Such approaches did not adequately consider the modality differences and exploit the modality correlations. We propose a novel approach for RGB-D face recognition that is able to learn complementary features from multiple modalities and common features between different modalities. Specifically, we introduce a joint loss taking activation from both modality-specific feature learning networks, and enforcing the features to be learned in a complementary way. We further extend the capability of this multi-modality (e.g., RGB-D vs. RGB-D) matcher into cross-modality (e.g., RGB vs. RGB-D) scenarios by learning a common feature transformation mapping different modalities into the same feature space. Experimental results on a number of public RGB-D face databases (e.g., EURECOM, VAP, IIIT-D, and BUAA), and a large RGB-D database we collected, show the impressive performance of the proposed approach.**

***Keywords*-RGB-D face recognition; complementary feature learning; common feature learning; joint loss; cross-modality**

I. INTRODUCTION

With the great advances of face recognition (FR) technologies, automatic FR systems have been used in many application scenarios, such as access control, ID de-duplication, and video surveillance. However, face recognition based on 2D face images is still facing the great challenges in unconstrained environment, such as variations of pose, illumination, expression, disguise, and plastic surgery. This is one of the main reasons why the state-of-the-art FR methods which reported nearly 99% accuracies on the LFW database (near frontal face images with illumination, expression, and occlusion variations) [1] have significant performance degradations in unconstrained scenarios.

Compared with 2D image based face recognition, 3D based face recognition is more robust to the above variations. However, 3D acquisition devices in the early years are expensive, and the speed of 3D image capturing is slow. With the development of depth sensing technology, such as the low-cost Kinect and RealSense RGB-D sensors, make it more convenient to obtain 2.5D face images in the modalities of RGB and depth (RGB-D). The price of RGB-D sensors also continue to drop, making it possible to integrate RGB-D sensors into commodity laptops, tablets, and even smartphones. As a result, face images of RGB-D modalities become more common, and require developing new method for RGB-D face recognition.

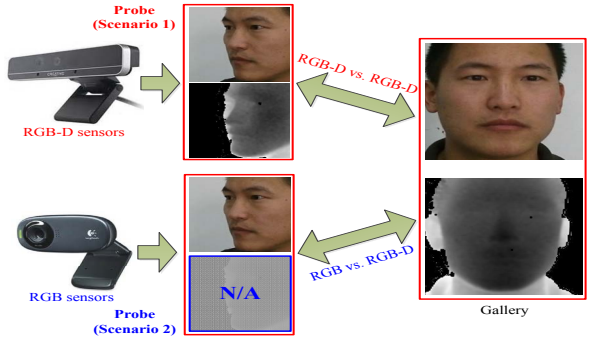

Figure 1. RGB-D face recognition (FR) consists of two typical senarios: (1) multi-modality matching, e.g., RGB-D probe vs. RGB-D gallery, where both the enrolled gallery and the probe images are captured using RGB-D sensors, and (2) cross-modality matching, e.g., RGB probe vs. RGB-D gallery, where the gallery images remain RGB-D, but the probe images are captured by RGB sensors. The proposed approach addresses the two problems by learning complementary and common features.

While there are a large number of methods reported for 2D (e.g., RGB) FR, studies on RGB-D FR are still very limited [2]. Early methods on RGB-D FR mainly used traditional image descriptors that are applied on RGB and depth modalities, independently. Then, either feature level or score level fusion was applied to fuse individual modalities [3]. However, the hand-crafted features show limitations in both representation effectiveness and generalization ability. In recent years, with the successful application of deep learning methods in many computer vision tasks, RGB-D FR also seeks to use deep learning methods for robust feature learning and classification [6]. However, existing deep learning based RGB-D FR methods either handle individual modalities using completely separate pipelines or treat all the modalities equally. These approaches did not adequately consider the different characteristics of individual modalities and exploit their correlations. Additionally, RGB-

*H. Han is the corresponding author.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

D FR is not a trivial problem dealing with only RGB-D vs. RGB-D FR matching. Instead, it contains two representative scenarios: (i) RGB-D vs. RGB-D (*multi-modality FR*) matching where both the probe and images are obtained using RGB-D sensors, and (ii) RGB vs. RGB-D (*cross-modality FR*) matching where the gallery images are enrolled in multi-modality but the probe images remain RGB [7]. Most of the existing RGB-D FR methods were limited to the former scenario, and ignore the latter scenario, which is very important to keep the compatibility with the existing 2D face databases.

Since RGB and depth describe face texture and 3D shape respectively, the two modalities should be correlated and complementary with each other. Through a complimentary learning, we expect to maximize the use of the complementary information, and obtain more discriminative features. In this work, we propose an RGB-D FR approach addressing both multi-modality and cross-modality FR problems. For multi-modality FR, we propose to jointly learn complementary features from multiple modalities. We start with two modality-specific features learning networks based on Inception-v2 [5], and introduce a joint loss architecture taking activation from both networks, to enforce the interferences between different modality-specific features learning networks. For cross-modality FR, we learn a common feature transformation to map different modalities into the same feature space, allowing measuring the similarity between a RGB and depth image. We split each RGB-D face image into a RGB and depth image, and get a training set with heterogeneous face images. We then train a heterogeneous feature learning network, and use it as the common feature transformation for all the modalities. With the proposed complementary feature learning and common feature learning, we are able to handle FR problems of RGB-D probe vs. RGB-D gallery, and RGB probe vs. RGB-D gallery.

Different from existing RGB-D feature learning methods [20], [21], our method uses a different architecture to learn more discriminative features from two modalities for multi-modality FR: (i) our shared layer allows interference between modalities at early layers, and (ii) our shared layer and each modality have their own losses, so that each modalitys discriminative feature can be retained.

The contributions of this work are two-fold: (i) a novel approach for complementary feature learning from multi-modality face images, leading to improved performance against unimodality face recognition; and (ii) a simple but efficient method for common feature learning between different modalities, allowing heterogeneous face recognition across modalities.

II. RELATED WORK

A. RGB-D Face Recognition

Compared with the large number of methods on 2D FR, studies on RGB-D face recognition are limited. Goswami et al. [2], [3] built a RGB-D face database (IIIT-D), and used it to study RGB-D FR. Their algorithm used concatenated Histogram of Gradient (HOG) features extracted from the entropy map and saliency map of each RGB-D image, and applied a Random Decision Forest (RDF) classifier for face matching. Face shape similarity based on landmark positions was also considered and fused with the RDF scores. The authors further extended their work by learning mapping and reconstruction between RGB and depth using a conventional deep autoencoder [6]. Hg et al. [8] built a RGB-D face database (VAP), and proposed a face detection algorithm based on curvature analysis, but no RGB-D FR method was reported in [8]. Min et al. [9] built a RGB-D face database (EURECOM), and evaluated the performance of several baseline methods like PCA, LBP, SIFT, etc. Zhang et al. [10] collected a relatively large RGB-D face database (BUAA Lock3DFace), and provided baseline results of the depth modality using the ICP algorithm. Hsu et al. [11] performed RGB-D FR by reconstructing 3D face from a RGB-D image. Xu et al. [12] used a distance metric learning algorithm for RGB-D FR by learning a common distance metric from RGB-D data. Most of these methods used traditional feature descriptors, and applied either feature concatenation fusion or score-level fusion.

There are also some methods on RGB-D based scene segmentation, object recognition, gesture recognition, and action recognition. Due to the limited space, we refer interested readers to literatures such as [13], [14], [15].

B. Deep RGB-D Feature Learning

With the successful applications of deep learning methods in computer vision tasks based on 2D images, RGB-D FR based on deep learning are drawing increased attentions. Wu et al. [16] proposed a 3D-ShapeNets for RGB-D object recognition and shape completion, which uses volumetric depth representation as input. Su et al. [17] used a multi-view deep model in object recognition. Most of these methods used only the shape information (either 3D shape or depth), but did not utilize the RGB information. Lee et al. [18] proposed to learn deep features from both RGB and enhanced depth images, and built a joint SVM classifier for face verification. Some work also studied the cross-modality face recognition problem [23], which may also present in RGB-D face recognition.

Recently, Socher et al. [19] proposed a Convolutional Recursive Neural Network (CRNN) for RGB-D object recognition, in which two CNN networks were separately trained using the RGB and depth images, and the learned CNN features were fed into two RNN networks to get the final features. Wang et al. [20] proposed a multi-modal sharable and modal-specific feature learning framework for RGB-D object recognition, and later they designed a multi-modal deep learning framework with a new loss to learn specific and correlative features [21]. Ngiam et al. [4] also

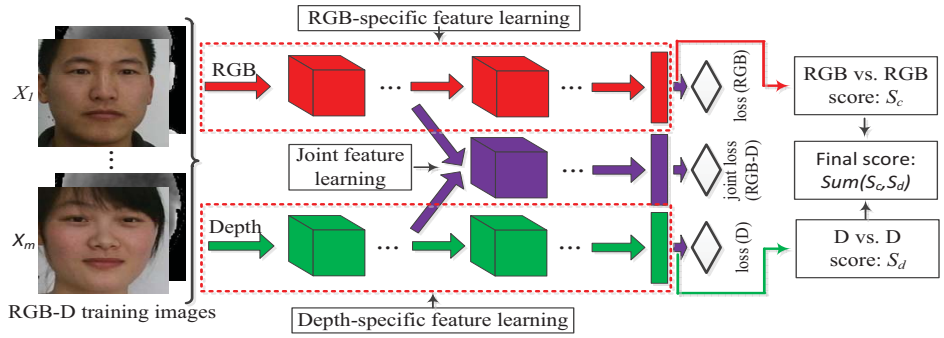

Figure 2. Overview of the proposed complementary feature learning approach from RGB and depth modalities, which handles multi-modality FR scenario such as RGB-D probe vs. RGB-D gallery.

proposed a novel application of deep networks to learn features over multiple modalities. Shahroudy et al. [15] proposed a deep autoencoder based shared-specific feature factorization network to separate input multimodal signals into a hierarchy of components, and used a structured sparsity learning machine for action recognition from RGB-D videos. There are also deep learning approaches on VIS-NIR (cross-modality) FR, e.g., [30], which used a shared layer to learn modality-invariant identity features; however, its performance for RGB-D FR is not known.

We can notice that most of the existing RGB-D feature learning methods either process individual modalities with completely separate pipelines or treat all the modalities equally using the same pipeline.

III. PROPOSED APPROACH

A. Overview

The proposed multi-modal face recognition approach consists of complementary feature learning (see Fig. 2), and common feature learning (see Fig. 6), addressing the two essential problems of multi-modality matching and cross-modality matching, respectively.

Formally, let $\{X_i\}_{i=1}^m$ denote the training set with m multi-modality face images of n subjects. For complementary feature learning with multi-modality face images, as shown in Fig. 2, we enforce the feature complementarity by simultaneously optimizing the per modality feature transformations and the joint feature transformation. The optimization can be formulated by minimizing

$$\begin{aligned}\underset{\{W^j\}_{j=1}^t, W^{\mathbb{J}}}{\arg\min} \sum_{j=1}^{t}\sum_{i=1}^{m} \mathcal{L}\big(y_i, \mathcal{F}(X_i^j, W^j)\big) \\ +\mathcal{L}\big(y_i, \mathcal{F}(X_i, W^{\mathbb{J}})\big)\end{aligned}, \tag{1}$$

where W^j denotes the modality-specific feature transformation (t modalities in total); $W^{\mathbb{J}}$ denotes the joint feature transformation applied to all the t modalities; $\mathcal{L}$ is a prescribed loss function, such as softmax loss; y_i denotes the identity of each face image. By introducing the second loss term, we expect that the feature learning for individual modalities could interfere with each other, leading to features with more complementarity.

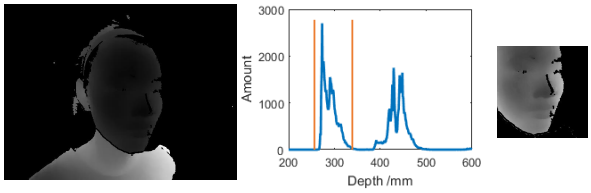

Figure 3. The processing of a depth face image: (a) original depth image, (b) depth histogram of (a) used to determine the truncation points of the face region, and (c) cropped depth face image after depth truncation.

For common feature learning across modalities, as illustrated in Fig. 6, we choose to optimize a common feature transformation W^{cd} that maps individual modalities into a common feature space. The optimization can be formulated by minimizing

$$\underset{W^{cd}}{\arg\min} \sum_{j=1}^{t}\sum_{i=1}^{m} \mathcal{L}\big(y_i, \mathcal{F}(X_i^j, W^{cd})\big). \tag{2}$$

This formulation differs the proposed approach from the widely used approaches that map individual modalities into a common feature space by learning a separate feature transformation per modality. The proposed approach only need to learn a uniform transformation, and thus has much lower computational cost.

We should point out that our formulations in (1) and (2) can generalize to face recognition scenarios with more than two modalities, such as RGB, depth, sketch, etc. In this work, we focus on face recognition with two modalities (RGB and depth).

B. Implementation Details

1) Depth image processing: We align all the RGB-D face images based on five facial landmarks (two eye centers, nose tip, and two mouth corners). The five landmarks are usually available for the RGB modality only, we map them to the depth image by using the camera's intrinsic parameters. The processing of the depth image plays an important role to the final FR performance. We propose a novel pipeline

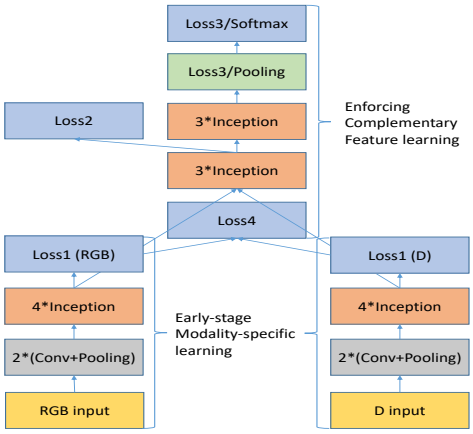

Figure 4. The network architecture of our complementary feature learning.

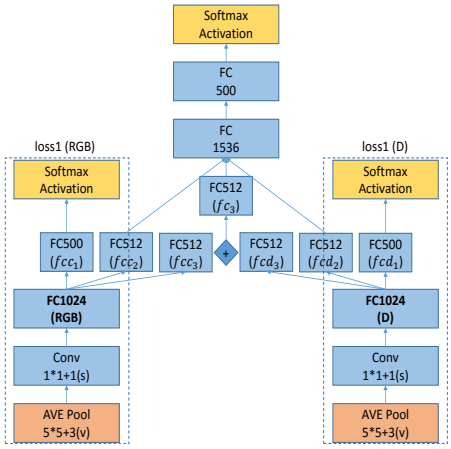

Figure 5. The details of the loss4 unit introduced to enforce complementary feature learning. The plus symbol denotes the element-wise average.

to process the depth images. We first compute the depth histogram from the original 16-bit depth face image. We then determine the truncation points as the first peak in the depth histogram. Specifically, we use a threshold of 50 pixels, for which a local extreme must be larger than to be a peak or smaller than to be a valley of the histogram. The two valleys on both sides of the 1st peak are used as the truncation points. Finally, we crop the face image after depth truncation. The process is illustrated in Fig. 3. We can see that the proposed processing pipeline retains better contrast of the face depth than the processed depth images by [3] (see Fig. 7 (b)).

2) Complementary Feature Learning: Following the success of convolutional neural networks (CNNs) in computer vision tasks like image classification, and face recognition (mainly based on RGB images), we choose to use a CNN architecture to simultaneously optimize the per modality feature transformations and the joint feature transformation in (1). Specifically, we start with two Inception-v2 [5] networks, handling the modality-specific feature transformations of RGB and depth, respectively. We introduce a new loss unit (namely loss4), which takes into account the activations from both modality-specific feature learning networks, and provides feedback on how the feature learning at early layers of the modality-specific networks should be performed in order to maximize their complementarity. The network architecture of the proposed complimentary feature learning is shown in Fig. 4.

Loss1 and loss2 have the similar structure as the loss in the auxiliary classifiers of [22], each containing an average pooling layer, a convolutional (Conv) layer, a fully connected (FC) layer, and finally a softmax classifier. The difference is that we insert a batch normalization (BN) layer after each Conv layer. Loss3 is a per modality loss, similar to loss1 and loss2, to retain per modality discriminative features. We design a FC layer (i.e., FC1536) containing three concatenated parts of activations. Two of them (i.e., fcc_2 and fcd_2) are connected to each of the per modality feature learning networks, through an average pooling layer, a Conv layer, and a FC layer. The third part (i.e., fc_3) is connected to both modality feature learning networks through an element-wise average operation. Each of the three parts consists of 512 nodes, leading to $1,536$ nodes in total. For the introduced loss4, its details are given in Fig. 5. The final layer of the loss4 architecture is a softmax classifier. Loss4 influences the training of the early layers (i.e., FC1024(RGB) and FC1024(D)) of the modality-specific feature learning networks through gradient backpropagation

$$\nabla fcc = \sum_{i=1}^{3} \nabla fcc_i, \nabla fcd = \sum_{i=1}^{3} \nabla fcd_i \tag{3}$$

Finally, we use the features of FC1024(RGB) and FC1024(D) (in Fig. 5) to calculate the RGB-to-RGB and depth-to-depth matching scores (S_c and S_d) via correlation, and fuse them using a weighed sum rule

$$S = \frac{p_1}{p_1 + p_2} S_c + \frac{p_2}{p_2 + p_1} S_d, \tag{4}$$

where p_1 and p_2 are the observed face matching accuracies by each modality alone.

While the modality-specific feature learning networks aim at attaining a discriminative representation for each of the RGB and depth modalities, the introduced loss4 unit enforces the interference between the modality-specific feature learning networks, leading to the final features with more complementarity.

3) Common Feature Learning: Most of the existing multi-modal FR methods assume a RGB-D vs. RGB-D scenario. However, in many situations, the depth modality may not be available. For example, millions of cameras used by access control and video surveillance remain RGB, which implies that the probe face images are RGB instead of RGB-D. Therefore, extending the compatibility of RGB-D face recognition system to handling RGB vs. RGB-D face matching is strongly required.

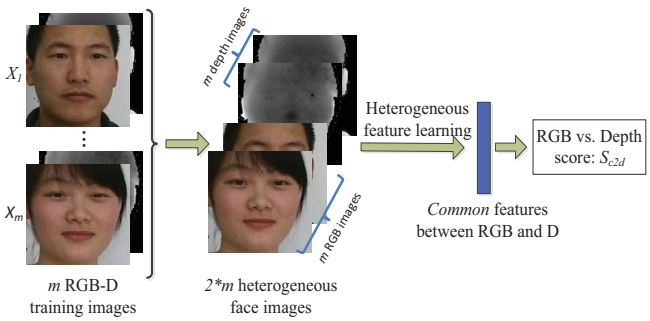

Figure 6. Overview of the proposed common feature learning approach between RGB and depth modalities, which handles cross-modality FR scenario such as RGB probe vs. depth gallery.

We propose to handle the cross-modality matching of RGB to RGB-D by dividing it into *two sub-problems*: (i) RGB to RGB matching, and (ii) RGB to depth matching.

RGB to RGB matching. For RGB to RGB matching, we can directly use the proposed complementary feature learning network above, and obtain a RGB-to-RGB matching score (S_{c2c}).

RGB to Depth matching. For RGB to Depth matching, we choose to learn common features between RGB and depth. Common feature learning between two different modalities is not new [23][24], but most of these methods used hand-crafted features followed by subspace learning.Additionally, these methods may require learning a feature transformation for each modality. We propose to learn a single common feature transformation that can be applied to individual modalities. Specifically, as shown in Fig. 6, given the m RGB-D training images $\{X_i\}_{i=1}^{m}$, we split each RGB-D image into a RGB image and a depth image, leading to $2m$ heterogeneous face images in total in the training set. We treat the optimization of the common feature transformation as a heterogeneous feature learning problem, and optimize a heterogeneous feature learning network, e.g., Inception-v2, from the $2m$ heterogeneous face images in the training set. The features from the FC layers in loss1 and loss2 can be used as the common feature representation for both RGB and depth modalities, and obtain a RGB-to-depth matching score (S_{c2d}). Finally, the RGB vs. RGB-D matching score can be calculated by a fusion of S_{c2c} and S_{c2d} using a weighted sum rule similar to (4).

IV. EXPERIMENTAL RESULTS

A. Databases and Experimental Settings

We provide evaluations on a number of widely used RGB-D databases including EURECOM [9], VAP [8], IIIT-D [2], and BUAA [10] RGB-D databases, as well as a RGB-D database we collected. **EURECOM** consists of 936 RGB-D images of 52 subjects, which were captured using Kinect I with face pose and expression variations. **VAP** contains 1,581 RGB-D images of 31 subjects, which were captured using Kinect I with 17 different poses and facial expressions. **IIIT-D** contains 4,605 RGB-D images of 106 subjects, which were captured using Kinect I with normal illumination and a few pose and expression variations. **BUAA** has much more subjects than the previous databases, and contains 5,711 RGB-D video sequences of 509 subjects, which were recorded using Kinect II with small variations in pose, expression, occlusion, and time. **Our RGB-D database** was captured using RealSense II instead of Kinect, containing about 845K RGB-D images of 747 subjects with continuous pose variations and a few illumination changes [1]. Example RGB-D images from these databases are shown in Fig. 7.

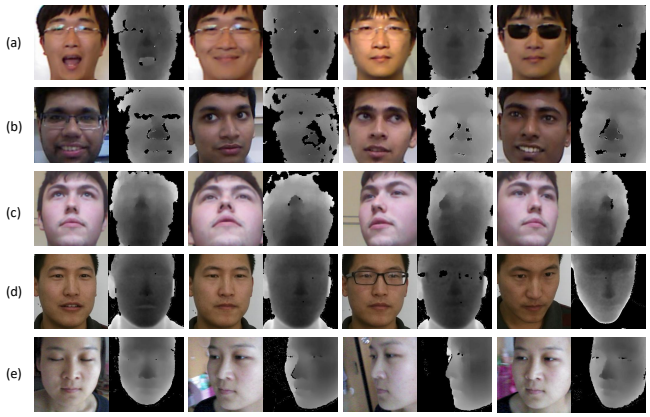

Figure 7. Example RGB-D face images from the (a) EURECOM, (b) IIIT-D, (c) VAP, (d) BUAA, and (e) our databases. We notice RealSense II captures more detailed depth information of face than Kinect II. Our preprocesing in Section III-B1 is applied on all the databases except for IIIT-D.

Given the above RGB-D databases, we provide evaluations of face identification from several perspectives: (1) multi-modality FR, (2) cross-modality FR, (3) multi-modal vs. unimodal FR, and (4) generalization ability evaluation. The testing protocols provided with the EURECOM, VAP, and BUAA databases only contain the gallery and probe sets. In addition, both EURECOM and VAP databases are very small, making it not possible to define a new training set. In this case, we train the network on BUAA, and directly use the learned model for feature extraction on EURECOM and VAP. For BUAA, since there are more than 500 subjects, we randomly select about 330 subjects for training. For the remaining 176 subjects, we use one frontal face image per subject as gallery, and the remaining face images as probe. Similarly, for our RGB-D database, we randomly select 500 subjects for training. For the remaining 247 subjects, we use one frontal face image per subject as gallery, and the remaining face images as probe. Both the RGB and depth face images are aligned based on the five facial landmarks, and normalized into the size of 256×256. For the RGB-D databases without provided facial landmarks, we use an open

[1]We plan to place our RGB-D dataset in the public domain upon the approval of our sponsors and volunteers.

Modality	GoogLeNet	AlexNet	VGG-16
RGB	89.9%	69.7%	71.4%
Depth	88.9%	76.9%	78.3%
Score fusion	96.8%	89.7%	90.4%
Proposed	**97.3%**	**90.5%**	**92.8%**

Table I
RANK-1 IDENTIFICATION ACCURACIES OF THE PROPOSED APPROACH WITH DIFFERENT BASIC NETWORKS (ALEXNET [25], GOOGLENET-BN [5], VGG-16 [26]) ON OUR RGB-D DATABASE SHOW THAT THE PROPOSED COMPLEMENTARY FEATURE LEARNING WITH DIFFERENT CNNS IS CONSISTENTLY HELPFUL FOR MULTI-MODAL FR.

source SeetaFaceEngine [2] for facial landmarks detection.

We use a batch size of 32 for training both the complementary and common feature learning networks. The base learning rate is set to 0.045 with a step update policy applied every $6K$ iterations. We use a step multiplier coefficient gamma of 0.9, weight decay of 0.0002, and momentum of 0.9. Some of the parameters are set heuristically; our method still achieved higher accuracies than the state of the art.

B. Influence of Different CNN Architectures

We used three different networks, i.e., AlexNet [25], GoogLeNet-BN [5], VGG-16 [26] to evaluate our complementary feature learning. All the networks were pre-trained using the CASIA WebFace database [27] for a classification task. For AlexNet and VGG-16, the joint loss was connected to the FC6 layers of the modality-specific feature learning networks. Performance on our RGB-D database by different networks are shown in Table I. We can see that the proposed approach is consistently helpful. Since GoogLeNet performs the best among the three networks, we choose to use GoogLeNet [5] as the basic network in our complementary and common feature learning.

C. Multi-modality Face Recognition

We evaluate the proposed approach for multi-modality FR on the EURECOM, VAP, IIIT-D, and BUAA databases, and the RGB-D dataset we collected. As a comparison, we also report the state-of-the-art performance by [3], [6], [28], [11] on these public-domain databases in the related publications. While each of the state-of-the-art methods only reported their performance on a single database, our results are reported on all the databases. The rank-1 identification accuracies by individual methods are given in Table II. The proposed approach outperforms [3] by a large margin, and achieves almost the same accuracy as the state-of-the-art method [6] on the IIIT-D database. However, as we pointed out in III-B1, the depth images released in IIIT-D was not preprocessed optimally, limiting the complementary feature learning by the proposed approach. On the EURECOM and VAP databases, which are relatively small, and do not have a training set, the proposed approach still outperforms the state-of-the-art method [28] on VAP by a large margin (7%).

[2]https://github.com/seetaface

Method	RGB-D vs. RGB-D face identification (rank-1)				
	IIIT-D	VAP	EURE	BUAA	Our DB
RISE [3]	86.0%	-	88.0%	-	-
AE [6]	**98.7%**	42.2%*	47.1%*	38.2%*	-
FLD [28]	-	83.1%	-	-	-
Reconst.[11]	-	-	91.2%	-	-
GoogLeNet 4-channel	97.6%	82.2%	96.0%	87.3%	91.5%
Score fusion	97.8%	88.6%	**96.3%**	90.1%	96.8%
Proposed	98.6%	**90.8%**	**96.3%**	**90.8%**	**97.3%**

Table II
COMPARISONS OF THE PROPOSED APPROACH AND THE STATE-OF-THE-ART METHODS [3], [6], [28], [11] FOR MULTI-MODALITY FR (I.E., RGB-D VS. RGB-D) ON THE PUBLIC-DOMAIN (IIT-D, VAP, EURECOM, AND BUAA) RGB-D DATABASES, AND OUR RGB-D DATABASE.*WE RE-IMPLEMENTED AE [6] BASED ON THE BEST OF OUR UNDERSTANDING.

These results demonstrate the effectiveness of the proposed approach in learning complementary features from multi-modality face images.

Besides the results reported by the state-of-the-art methods, we use two additional baseline methods: (i) GoogLeNet trained with a four-channel input (R, G, B, and D), where the depth is treated equally as the color channels, and (ii) score level fusion of two GoogLeNets; each is trained for face matching with one modality. Again, the proposed complementary feature learning approach outperforms both baseline methods on all the databases except for EURECOM. The advantages of the proposed approach become more evident on the large scale RGB-D databases such as BUAA and our dataset. These results indicate that low-level fusion of multi-modalities (i.e., 4-channel GoogLeNet) is not an effective way to explore the complementarity between different modalities. This is understandable because the RGB and depth modalities are completely heterogeneous; while RGB describes the face texture, depth reflects the face shape. Score level fusion of RGB and depth is also not optimum because the two networks are learned in an isolated way. Examples of some correct and incorrect FR results by the proposed approach are shown in Figure 8.

D. Cross-modality Face Recognition

Cross-modality FR, e.g., RGB probe vs. RGB-D gallery, has wide potential applications, we evaluate the proposed approach under this scenario using three relatively large RGB-D databases (IIIT-D, BUAA, and our dataset). Since there is not known cross-modality FR results reported in the state-of-the-art methods such as [3], [6], [28], [11], we directly analyze the accuracy of the proposed common feature learning. Since the RGB vs. RGB-D matching problem is divided into RGB vs. RGB and RGB vs. depth matching, we also report the accuracy for each step. As shown in Table III, the proposed approach achieves 65.3%, 57.8%, and 50.0% rank-1 accuracies for RGB vs. depth matching on IIIT-D, BUAA, and our database, respectively. Although

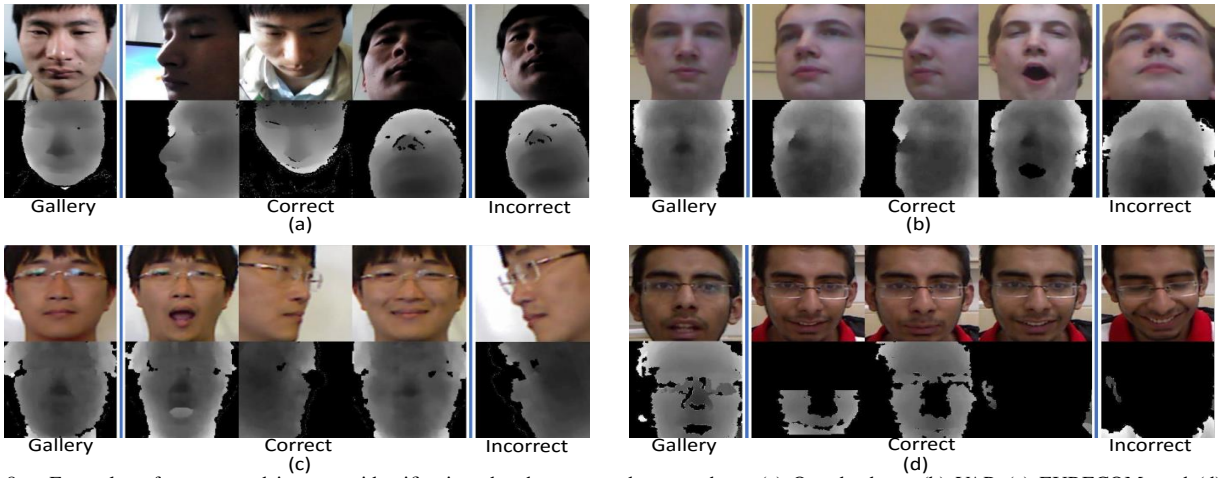

Figure 8. Examples of correct and incorrect identifications by the proposed approach on (a) Our database, (b) VAP, (c) EURECOM, and (d) IIIT-D RGB-D databases. The gallery images are RGB-D; 'correct' means correct identifications using RGB-D, but incorrect identifications using only RGB; 'incorrect' means incorrect identifications using RGB-D.

Modalities of probe (P) and gallery (G)	IIIT-D	BUAA	Our DB
RGB (P) vs. RGB (G)	**99.0%**	90.4%	89.9%
RGB (P) vs. Depth (G)	65.3%	57.8%	50.0%
RGB (P) vs. RGB-D (G)	98.7%	**90.7%**	**92.7%**

Table III
EFFECTIVENESS OF THE PROPOSED APPROACH FOR CROSS-MODALITY FR ON THE IIIT-D, BUAA, AND OUR RGB-D DATABASES REPORTED IN TERMS OF THE RANK-1 IDENTIFICATION ACCURACY.

these accuracies are much lower than the RGB vs. RGB accuracies, such results are promising given the challenging RGB vs. depth matching scenario. The score level fusion of RGB vs. RGB and RGB vs. depth matching scores leads to improved accuracies than using RGB modality alone on BUAA and our database. This experiment suggests that for cross-modality FR such as RGB probe vs. RGB-D gallery, the depth information in the gallery should not be simply discarded. Instead, we can use common feature learning to explore the supplemental information and enhance the capability of the FR systems. Additionally, cross-modality FR also allows multi-modality FR systems to be compatible with the huge amount of existing 2D face databases.

E. Multi-modality vs. Unimodality FR

From the above sections, we may recognize the advantages of multi-modality FR against unimodality FR. But the accuracies for unimodality FR with RGB images is only reported based on GoogLeNet. We would like to provide more comparisons by using a state-of-the-art RGB FR method [29] as the baseline. Similar comparisons between multi-modality and unimodality FR performance were provided in [6], which also show that multi-modality is more robust than unimodality FR. But the RGB FR methods used for comparisons in [6] were limited to hand-crafted features.

The state-of-the-art RGB FR method, VIPLFace [29], reported about 99% accuracy on the public-domain LFW database [1], and should be a strong baseline of RGB FR. We fine-tune the VIPLFace FR model on the training sets of BUAA and our RGB-D database, respectively, and achieve 92.5% and 94.2% rank-1 accuracies on the corresponding testing sets, which are higher than the baseline accuracies (90.4% and 89.9%) by GoogLeNet (see Table IV). However, after the VIPLFace RGB model is fused with the GoogLeNet depth model, the accuracy can be improved to 93.2% and 97.4% on BUAA and our RGB-D database, respectively.[3] A t-test gives a p-value of 0.08, implying the improvement against using RGB alone is significant. This suggests that the complementary information in the depth modality is helpful for the state-of-the-art RGB face matchers.

Similarly, fusion of the RGB and depth models by GoogLeNet also leads to higher accuracy than using RGB alone. Among the published methods, unimodality FR on the BUAA database was reported in [10] using only the depth image. Although their testing protocol is different, we still provide their result as a reference. The average accuracy of [10] is 34.5% for depth based FR under four scenarios such as expression, pose, occlusion, and aging. Our accuracy for depth based FR on BUAA is 66.0%, which is much better than the state-of-the-art result on BUAA. The experimental results suggest that multi-modality FR could be a good choice in the deployment of FR systems in future applications, particularly considering the decreasing prices of commodity RGB-D sensors.

V. CONCLUSIONS

In this paper, we introduce a RGB-D face recognition approach consisting of complementary feature learning and common feature learning, which are effective for multi-modality face recognition and cross-modality face recognition scenarios. In complementary feature learning, a novel

[3]We also fine-tuned VIPLFace RGB FR model using the depth images in our RGB-D database, but we notice the performance is poor. This is interpretable because VIPLFace was NOT pre-trained using depth images, but $470K$ RGB images from CASIA WebFace [29].

Modalities of probe (P) and gallery (G)	BUAA	Our DB
S_1: RGB (P) vs. RGB (G)-VIPLFace [29]	92.5%	94.2%
S_2: RGB (P) vs. RGB (G)-GoogLeNet	90.4%	89.9%
S_3: Depth (P) vs. Depth (G)-GoogLeNet	66.0%	88.9%
RGB-D (P) vs. RGB-D (G) – S_1 + S_3	**93.2%**	**97.4%**
RGB-D (P) vs. RGB-D (G) – S_2 + S_3	90.1%	96.8%
RGB-D (P) vs. RGB-D (G) – Proposed	90.8%	97.3%

Table IV
COMPARISONS OF RANK-1 IDENTIFICATION ACCURACY BETWEEN OUR MULTI-MODAL FR METHODS AND A STATE-OF-THE-ART RGB FR METHOD ON BUAA AND OUR RGB-D DATABASES.

loss is designed to take into account both the RGB and depth features, enforcing the interference between the modality-specific feature learning networks. Common feature learning resolves cross-modality matching problem by learning a common feature transformation for all the modalities, and transform individual modality into the same feature space. Experiments are provided under a number of practical scenarios including multi-modality face recognition, cross-modality face recognition, and multi-modality vs. unimodality face recognition. The results show that bottom-level or top-level fusion of RGB and depth may not make the best use of their complementary information. In addition, while cross-modality FR is very difficult and the accuracy is much lower than RGB vs. RGB FR, it still learns complementary information, and improves RGB vs. RGB FR accuracy.

ACKNOWLEDGMENT

This research was supported in part by the National Basic Research Program of China (grant 2015CB351802), Natural Science Foundation of China (grants 61732004, 61672496, and 61390511), External Cooperation Program of CAS (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), and Youth Innovation Promotion Association CAS (grant 2018135).

REFERENCES

[1] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[2] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On RGB-D face recognition using Kinect. In *BTAS*, pages 1–6, 2013.

[3] G. Goswami, M. Vatsa, and R. Singh. RGB-D face recognition with texture and attribute features. *IEEE Trans. IFS*, 9(10):1629–1640, 2014.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.

[6] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa. RGB-D face recognition via learning-based reconstruction. In *IEEE BTAS*, pages 1–7, 2016.

[7] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen. Improving 2D face recognition via discriminative face depth estimation. In *ICB*, pages 1–8, 2018.

[8] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B Moeslund, and G. Tranchet. An RGB-D database using Microsoft's Kinect for windows for face detection. In *SITIS*, pages 42–46, 2012.

[9] R. Min, N. Kose, and J. Dugelay. KinectFaceDB: A kinect database for face recognition. *IEEE Trans. SMC: Systems*, 44(11):1534–1548, 2014.

[10] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3DFace: A large-scale database of low-cost kinect 3d faces. In *IEEE ICB*, pages 1–8, 2016.

[11] G. J. Hsu, Y. L. Liu, H. C. Peng, and P. X. Wu. RGB-D-based face reconstruction and recognition. *IEEE Trans. IFS*, 9(12):2110–2118, 2014.

[12] X. Xu, W. Li, and D. Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Trans. NNLS*, 26(12):3150–3162, 2015.

[13] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *IEEE IROS*, pages 821–826, 2011.

[14] X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: Features and algorithms. In *IEEE CVPR*, pages 2759–2766, 2012.

[15] A. Shahroudy, T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE Trans. PAMI*, 2017 (To appear).

[16] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE CVPR*, pages 1912–1920, 2015.

[17] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *IEEE ICCV*, pages 945–953, 2015.

[18] Y. Lee, J. Chen, C. Tseng, and S. Lai. Accurate and robust face recognition from RGB-D images with a deep learning approach. In *BMVC*, pages 123.1–123.14, 2016.

[19] R. Socher, B. Huval, B. Bath, C. D Manning, and A. Y Ng. Convolutional-recursive deep learning for 3d object classification. In *NIPS*, pages 665–673, 2012.

[20] A. Wang, J. Cai, J. Lu, and T. Cham. MMSS: Multi-modal sharable and specific feature learning for RGB-D object recognition. In *IEEE ICCV*, pages 1125–1133, 2015.

[21] A. Wang, J. Lu, J. Cai, T. Cham, and G. Wang. Large-margin multi-modal deep learning for RGB-D object recognition. *IEEE Trans. MM*, 17(11):1887–1898, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, pages 1–9, 2015.

[23] H. Han, B. F Klare, K. Bonnen, and A. K Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Trans. IFS*, 8(1):191–204, 2013.

[24] S. Klum, H. Han, B. Klare, and A. K Jain. The FaceSketchID System: Matching Facial Composites to Mugshots. *IEEE Trans. IFS*, 9(12):2248–2263, 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2015.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *ArXiv*, 2014.

[28] R. Bormann, T. Zwolfer, J. Fischer, J. Hampp, and M. Hagele. Person recognition for service robotics applications. In *IEEE RAS*, pages 260–267, 2013.

[29] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen. VIPLFaceNet: An open source deep face recognition sdk. *Frontiers of Computer Science*, 2016.

[30] R. He, X. Wu, Z. Sun, T. Tan. Learning invariant deep representation for NIR-VIS face recognition. In *AAAI*, pages 1097–1105, 2012.